RETRIEVAL METHOD USING SYNTAX INFORMATION AND SYSTEM THEREFOR

Publication number: JP2000076274

Publication date: 2000-03-14

Inventor: WATANABE HIDEO

Applicant: IBM

Classification:

- International: ***G06F17/27; G06F17/30;*** **G06F17/27; G06F17/30;** (IPC1-7): G06F17/30; G06F17/27

- European: G06F17/30T2P4N

Application number: JP19980245050 19980831

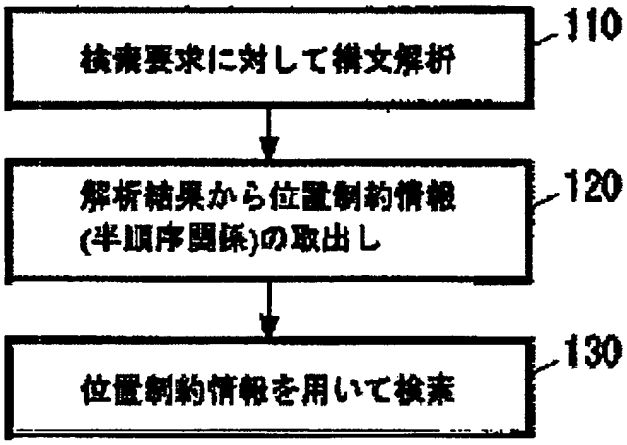
Priority number(s): JP19980245050 19980831

Also published as:

US6219664 (B1)

Abstract of **JP2000076274**

PROBLEM TO BE SOLVED: To provide retrieval method/system where the balance of the precision and the speed of syntax analysis is kept. SOLUTION: A retrieval request sentence is syntax-analyzed and position restriction information of a keyword and a function word is taken out from a syntax analysis result as a semi-order relation. The sentence satisfying the semi-order relation is retrieved from the document of a retrieval object without syntax-analyzing the document of the retrieval object. At that time, the sentence whose shorter length of a context satisfying the semi-order relation is retrieved than longer sentence as the sentence of high similarity for retrieving the sentence satisfying the semi-order relation from the document of the retrieval object.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開2000－76274
（P2000－76274A）
(43)公開日　平成12年３月14日(2000.3.14)

(51)Int.Cl.'	識別記号	FI		テーマコード'(参考)
G06F 17/30		G06F 15/403	330C	5B075
17/27		15/38	J	5B091
		15/40	370A	
		15/403	350C	

審査請求　未請求　請求項の数７　ＯＬ　（全５頁）

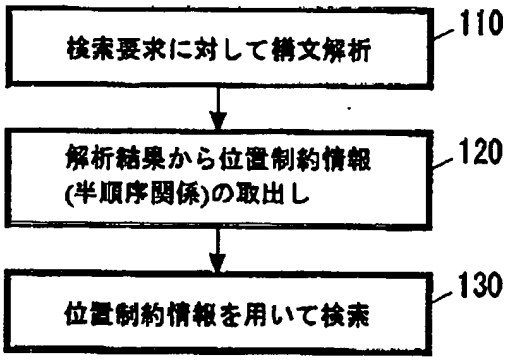
(21)出願番号　特願平10－245050
(22)出願日　平成10年８月31日(1998.8.31)

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレイション
INTERNATIONAL BUSINESS MASCHINES CORPORATION
アメリカ合衆国10504、ニューヨーク州　アーモンク　（番地なし）
(74)代理人　100086243
弁理士　坂口　博　（外１名）

最終頁に続く

(54)【発明の名称】　構文情報を用いた検索方法およびシステム

(57)【要約】
【課題】構文解析の精度とスピードのバランスのとれた検索方法及びシステムを提供することである。
【解決手段】検索要求文を構文解析し、構文解析結果から、キーワードと機能語（FNWORD）の位置制約情報を半順序関係として取り出す。そして検索対象文書を構文解析をすることなく、それらの半順序関係を満たす文を検索対象文書から検索する。またこの時、検索対象文書から半順序関係を満たす文を検索するにあたり、半順序関係を満たす文脈の長さが短い文を、より類似度の高い文として検索する。



1

【特許請求の範囲】

【請求項１】検索要求文を検索対象文書から検索する、検索システムであって、、（１）検索要求文を構文解析する手段と、（２）前記構文解析結果から、位置制約情報を半順序関係として取り出す手段と、（３）検索対象文書を構文解析をすることなく、前記半順序関係を満たす文を検索対象文書から検索する手段と、を具備することを特徴とする、検索システム。

【請求項２】前記検索する手段（３）が、前記検索対象文書から前記半順序関係を満たす文を検索するにあたり、前記半順序関係を満たす文脈の長さが短い文を、より類似度の高い文として検索する手段である、請求項１記載のシステム。

【請求項３】検索要求文を検索対象文書から検索する、ネットワーク上の検索システムであって、（１）検索要求文をネットワークを通じて受信する手段と、（２）検索要求文を構文解析する手段と、（３）前記構文解析結果から、位置制約情報を半順序関係として取り出す手段と、（４）検索対象文書を構文解析をすることなく、前記半順序関係を満たす文を検索対象文書から検索する手段と、（５）検索結果を送信する手段と、を具備することを特徴とする、検索システム。

【請求項４】前記検索する手段（４）が、前記検索対象文書から前記半順序関係を満たす文を検索するにあたり、前記半順序関係を満たす文脈の長さが短い文を、より類似度の高い文として検索する手段である、請求項３記載のシステム。

【請求項５】検索要求文を検索対象文書から検索する、検索方法であって、、（１）検索要求文を構文解析する段階と、（２）前記構文解析結果から、位置制約情報を半順序関係として取り出す段階と、（３）検索対象文書を構文解析をすることなく、前記半順序関係を満たす文を検索対象文書から検索する段階と、を有することを特徴とする、検索方法。

【請求項６】前記検索する手段（３）が、前記検索対象文書から前記半順序関係を満たす文を検索するにあたり、前記半順序関係を満たす文脈の長さが短い文を、より類似度の高い文として検索する段階である、請求項５記載の方法。

【請求項７】検索要求文を検索対象文書から検索するプログラムを含む媒体であって、該プログラムが、（１）検索要求文を構文解析する機能と、（２）前記構文解析結果から、位置制約情報を半順序関係として取り出す機能と、（３）検索対象文書を構文解析をすることなく、前記半順序関係を満たす文を検索対象文書から検索する機能と、を具備することを特徴とする、プログラムを含む媒体。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本願は、構文情報を活用した検索方法およびそのシステムに関し、特に検索要求に対してだけ構文解析を行い文書検索を行う方法およびそのシステムに関する。

2

【０００２】

【従来の技術】現在使われているワールドワイドウェブ（WWW）上の検索システムは、キーワード型か全文検索型のどちらかが使われているのが普通であり、この様なシステムでは検索結果が非常に大量に提示され、目指すドキュメントにたどり着くまでに大変な苦労をするという問題がある。このような問題に対処するために様々な試みがされてきた。その一つが、検索要求を幾つかのキーワードの論理積や論理和ではなく文章にし、この検索要求文章と似たものを検索するというものである。この方式は技術的に以下の方式に分類することができる。

【０００３】

(1) ベクタースペースモデル

(2) キーワード位置制約型マッチング方式

(3) 構文マッチング方式

【０００４】(1)のベクタースペースモデル（Salton, G., " Automatic Text Processing: the transformation, analysis, and retrieval of information by computer," Addison-Wesley Publishing, 1989.）は、文書と検索要求それぞれをキーワードを軸としたベクターとみなし、そのベクター間の距離により類似度を計算する方式である。しかし、この方式は結局検索要求中のキーワードを単に独立に出現したと考えているため、大きな文書の中にたまたま検索要求中のキーワードが含まれていたというような場合に対処できないという欠点がある。(2)のキーワード位置制約型マッチング方式（田中英輝、「長い日本語表現の高速類似検索手法」、情報処理学会言語処理研究会資料 NLWG121-10, 1997）とは、検索要求からキーワードを取り出し、それらキーワードの出現位置に関する全順序関係を満たすものをマッチするとするものである。この方式は(1)よりも良いが、やはりキーワード間の出現位置だけを制約にしている点で(3)に劣る。(3)は、検索要求と文書をともに構文解析し、構文木レベルでのマッチングを取る方式である。この手法は理想形であるが、残念ながら構文解析の精度とスピードの問題があり、広く普及するに至っていない。

【０００５】

【発明が解決しようとする課題】従って、本発明が解決しようとする課題は、構文解析の精度とスピードのバランスのとれた検索方法及びシステムを提供することである。また別の課題は、ネットワーク上での検索を効率よく行う方法及びシステムを提供することである。また別の課題は、検索対象文書を構文解析しない、検索方法及びシステムを提供することである。また別の課題は、検索要求文の位置制約情報を用いて検索を行う方法及びシステムを提供することである。

【０００６】

【課題を解決するための手段】上記課題を解決するために、まず検索要求文を構文解析し、構文解析結果から、キーワードと機能語（FNWORD）の位置制約情報を半順序関係として取り出す。そして検索対象文書を構文解析をすることなく、それらの半順序関係を満たす文を検索対象文書から検索する。またこの時、検索対象文書から半順序関係を満たす文を検索するにあたり、半順序関係を満たす文脈の長さが短い文を、より類似度の高い文として検索する。このような構成することにより、(2)のキーワード位置制約型マッチング方式に比べて構文解析を行うことによりより詳細な位置制約を抽出できる。また、(3)の構文マッチング方式に比べて検索対象文書を構文解析をしないため構文解析の不完全さに起因するマッチング精度の悪さとスピードの遅さという問題点を回避できる。

【０００７】図１に本発明の検索方法の基本フローチャートを示す。ますステップ１１０で検索要求文を構文解析する。次にブロック１２０で解析結果から得られる位置制約情報（半順序関係）を取り出す。そして最後にステップ１３０で位置制約情報（半順序関係）にマッチングする文を、検索対象文書から検索する。

【０００８】検索要求文の構文解析をより詳細に説明する。 検索要求文をQSとすると、その構文解析木QTは一般に以下のように表現できる。

【数１】

```
QT=TREE
TREE=(WORD)|(CHILD+ HEAD CHILD*)|(CHILD* HEAD CHILD+)
HEAD=('HEAD' TREE)
CHILD=(FUNC TREE)|(TREE FUNC)
FUNC='FN' WORD
```

ここで、FUNCはHEADとTREEの間の係り受けの関係を表す。以下に、検索要求の解析木の例を示す。

【数２】

```
「ＸＸＸ社のＹＹＹ社への提訴」
(((ＸＸＸ社) FN の)((ＹＹＹ社) FN への) (HEAD (提訴)))

"lawsuit of XXXCo. to YYYCo."
(((HEAD(lawsuit)) (FN of (XXXCo.)) (FN to (YYYCo.)))
```

上記の様な構文解析木からHEADとCHILDの間の位置情報を位置制約情報として用いる。取り出す位置制約情報は以下の様になる。

【０００９】・順序制約 ... CHILDとHEADはその位置関係を保持しなければならない。例えば、CHILDの後方にHEADがあることは、CHILD ＝＞ HEADと記述する。

・近傍順序制約 ... NODEのHEADワードとFNワードはその位置関係を保持するとともに近傍になければならない。ただし、近傍とはパラメータとして与えられる数値の語数以内にあることである。例えば、NODEの後方近傍にFNWORDがあることは、NODE －＞ FNWORD と記述する。

【００１０】よって、上記の日本語文の例からは以下の様な位置制約情報が得られる。

【数３】

```
ＸＸＸ社 ➔ 提訴
ＹＹＹ社 ➔ 提訴
ＸＸＸ社 ー の
ＹＹＹ社 ー への
```

また、英語文の例からは以下の様な位置制約情報が得られる。

【数４】

```
lawsuit ➡ XXXCo.
lawsuit ➡ YYYCo.
of → XXXCo.
to → YYYCo.
```

これらの位置制約情報を検索に使うことになる。ただし、その際に、一文、二文、一段落とこれらの制約を満たす文脈が小さい方がマッチングの類似度が高くなるようにする。

【００１１】従来技術(1)のベクタースペースモデルに比べると、キーワードの位置制約を用いている点で(2)のキーワード位置制約型マッチング方式と同様にすぐれているのは明らかである。また、(3)の構文マッチング方式に比べると、検索対象文書を構文解析しないことにより、構文解析の不完全さと構文木どうしのマッチングのスピードの遅さの問題がない点で優れている。(2)のキーワード位置制約型マッチング方式に比べると、構文木の依存関係から選られる位置制約により、より柔軟な検索が行える。例えば、以下のように検索要求中にＡ、Ｂ、Ｃ、Ｄ，Ｅ，Ｆという６つのキーワードがこの順番で存在し、以下の様な構文木を形成するものである場合、

【数５】

```
検索要求： A ... B ... C ... D ...E ... F
構文木：   ((FN fn1 ((FN fn2 (A)) (HEAD (B))))
           (FN fn3 ((FN fn4 (C) (FN fn5(D)) (HEAD (E))))
           (HEAD (F)))
文書１：   ... A ... B ... C ... D ... E ...F ...
文書２：   ... A ... B ... C ... D ... E ... F ...
文書３：   ... C ... D ... E ... A ... B ... F ...
文書４：   ... D ... C ... E ... A ... B ... F ...
```

(2)のキーワード位置制約型マッチング方式では文書１としかマッチできないが、本発明の手法では文書１から４まですべてのバリエーションにマッチ可能である。本手法では、あるHEAD語に係っている要素が複数ある場合にそれらが任意の順番で存在することを許したマッチングになっている。すなわち、(2)のキーワード位置制約型マッチング方式の手法はキーワードの位置制約を全順序関係として捉えているが、本手法では構文構造から得られる半順序関係として捉えている。さらに、本発明で

は機能語を用いていることにより、(2)のキーワード位置制約型マッチング方式に比べて絞り込みが可能である。従って上記の点を考慮することにより、(2)のキーワード位置制約型マッチング方式よりも高精度な検索が可能である。

【００１２】

【発明の実施の形態】本発明の方法をネットワーク上での検索システムに応用した実施例を説明する。特にインターネット上での検索では、検索結果が非常に大量に提示されるが本発明のネットワーク上での検索システムは、構文解析の精度とスピードのバランスのとれた検索が可能である。図４に本発明のネットワーク上での検索システムの処理の流れを示す。まずステップ４１０で検索要求文をネットワークを通じて受信する。次にステップ４２０で該検索要求文を構文解析する。構文解析で得られた構文解析木から位置制約情報を取り出す。この位置制約情報は、図２で示されるようにまずＨＥＡＤ，ＣＨＩＬＤの順序制約の取り出し，およびＨＥＡＤ，ＦＮＷＯＲＤの近傍順序制約の取り出しから構成される。これらを半順序関係と呼ぶ。次に処理は図４のステップ４４０移り、得られた位置制約情報（半順序関係）を用いて検索対象文書データベース４５０から検索する。このとき、検索対象文書から半順序関係を満たす文を検索するにあたり、半順序関係を満たす文脈の長さが短い文を、より類似度の高い文として検索するようにする。そしてステップ４６０で検索結果を検索要求元へ送信する。なお検索結果をこのとき表示するようにしてもよい。

【００１３】図３に本発明において使用される検索システムのハードウェア構成例を示す。システム１００は、中央処理装置（ＣＰＵ）１とメモリ４とを含んでいる。ＣＰＵ１とメモリ４は、バス２を介して、補助記憶装置としてのハードディスク装置１３（またはMO、ＣＤ－ＲＯＭ２３、ＤＶＤ等の記憶媒体駆動装置）とＩＤＥコントローラ２５を介して接続してある。同様にＣＰＵ１とメモリ４は、バス２を介して、補助記憶装置としてのハードディスク装置３０（またはMO２８、ＣＤ－ＲＯＭ２３、ＤＶＤ等の記憶媒体駆動装置）とＳＣＳＩコントローラ２７を介して接続してある。フロッピーディスク装置２０はフロッピーディスクコントローラ１９を介してバス２へ接続されている。

【００１４】フロッピーディスク装置２０には、フロッピーディスクが挿入され、このフロッピーディスク等やハードディスク装置１３（またはMO、ＣＤ－ＲＯＭ、ＤＶＤ等の記憶媒体）、ＲＯＭ１４には、オペレーティングシステムと協働してＣＰＵ等に命令を与え、本発明を実施するためのコンピュータ・プログラムのコード若しくはデータを記録することができ、メモリ４にロードされることによって実行される。このコンピュータ・プログラム（ＯＳ，検索プログラムなど）のコードは圧縮し、または、複数に分割して、複数の媒体に記録することもできる。

【００１５】システム１００は更に、ユーザ・インターフェース・ハードウェアを備え、入力をするためのポインティング・デバイス（マウス、ジョイスティック等）７またはキーボード６や、検索要求文、検索結果データ等をユーザに提示するためのディスプレイ１２を有することができる。また、パラレルポート１６を介してプリンタを接続することや、シリアルポート１５を介してモデムを接続することが可能である。このシステム１００は、シリアルポート１５およびモデムまたは通信アダプタ１８(イーサネットやトークンリング・カード)等を介してネットワークに接続し、他のコンピュータ等と通信を行う。好ましくは通信アダプタ１８を介して、検索要求文を受け取り、検索結果を該アダプタから送信する。またシリアルポート１５若しくはパラレルポート１６に、遠隔送受信機器を接続して、赤外線若しくは電波によりデータの送受信を行うことも可能である。

【００１６】スピーカ２３は、オーディオ・コントローラ２１によってＤ／Ａ（デジタル／アナログ変換）変換された音声信号を、アンプ２２を介して受領し、音声として出力する。また、オーディオ・コントローラ２１は、マイクロフォン２４から受領した音声情報をＡ／Ｄ（アナログ／デジタル）変換し、システム外部の音声情報をシステムにとり込むことを可能にしている。

【００１７】このように、本発明の検索システムは、通常のパーソナルコンピュータ（ＰＣ）やワークステーション、ノートブックＰＣ、パームトップＰＣ、ネットワークコンピュータ、コンピュータを内蔵したテレビ等の各種家電製品、通信機能を有するゲーム機、電話、ＦＡＸ、携帯電話、ＰＨＳ、電子手帳、等を含む通信機能有する通信端末、または、これらの組合せによって実施可能であることを容易に理解できるであろう。ただし、これらの構成要素は例示であり、その全ての構成要素が本発明の必須の構成要素となるわけではない。

【００１８】

【発明の効果】本発明により、従来の検索手法で実現困難であった構文情報を反映した検索が可能となる。全て構文木でマッチするというフルに構文解析を使用した手法に比べて十分高速かつ、単なるキーワードの位置情報制約を使う手法に比べてより詳細なマッチングが可能となる。さらに大量の検索結果が出てしまう現状のインターネット検索の問題点に対して、速度とスピードと精度の観点でバランスの取れた検索手法を提供できる。

【００１９】

【図面の簡単な説明】

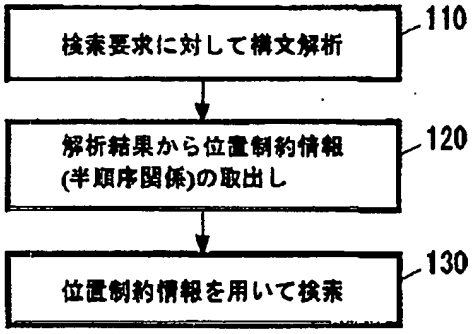
【図１】本発明の検索方法の基本フローチャートである。

【図２】位置制約情報における半順序関係の取り出しを示す図である。

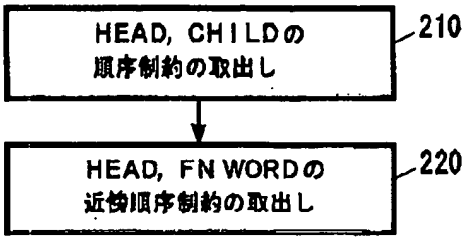
【図３】本発明において使用される検索システムのハードウェア構成例である。

＊【図４】本発明のネットワーク上での検索システムのフローチャートである。

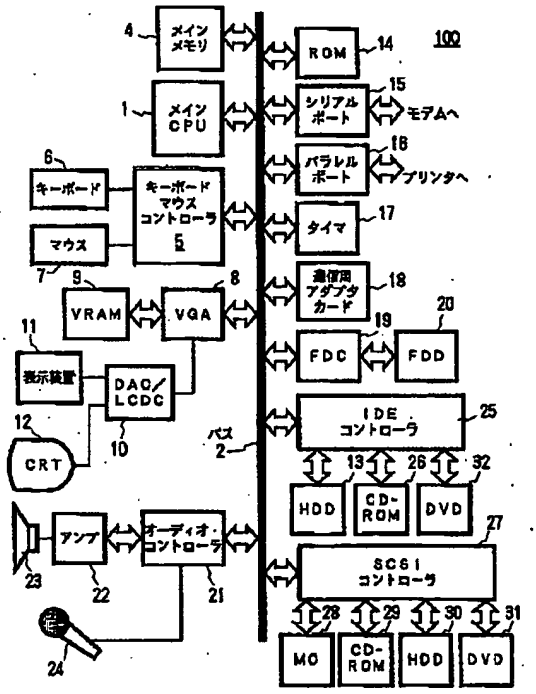
【図１】



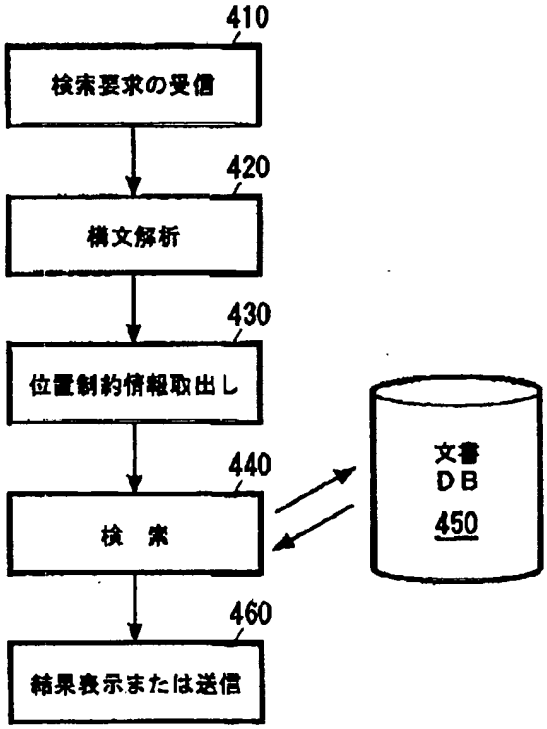
【図２】



【図３】



【図４】



フロントページの続き

(72)発明者　渡辺　日出雄
神奈川県大和市下鶴間1623番地14　日本アイ・ビー・エム株式会社　東京基礎研究所内

Ｆターム(参考)　5B075 ND03 PP24 PR10 QM05 QM08 UU06
5B091 AA15 CA05 CD03